What is Claimed Is

[c1] A method for automatic triage of a text passage outputted by an optical character recognition system, the OCR-output text passage having at least one OCR-output character, the method comprising:
determining at least one OCR-output character attribute for each OCR-output character;
determining an error rate for the OCR-output text passage using a triage model and the determined at least one OCR-output character attribute; and
comparing the determined error rate for the OCR-output text passage with an OCR-output text passage threshold error rate to perform an OCR-output text passage triage decision.

[c2] The method of claim 1, wherein determining an error rate for the OCR-output text passage comprises:
providing the at least one OCR-output character attribute to the triage model;
determining a character interpretation error value for each OCR-output character based on a probability of the at least one OCR-output character attribute being erroneously interpreted by the system; and
determining a text passage error value based on the at least one character interpretation error value determined for each OCR-output character.

[c3] The method of claim 2, further comprising:
determining a number representing a sum of OCR-output characters in the OCR-output text passage; and
dividing the text passage error value by the number representing the sum of OCR-output characters.

[c4] The method of claim 1, wherein determining at least one OCR-output character attribute for each OCR-output character comprises selecting the at least one OCR-output character attribute from a plurality of OCR-output character attributes.

[c5] The method of claim 4, wherein the plurality of OCR-output character attributes includes at least one of a character class, a confidence descriptor class, a language of the text passage, a text passage publication date, a typeface in

which the text passage is printed, an image-based feature of an individual character image and metadata attached to the text passage.

[c6] The method of claim 1, wherein the text passage to be triaged includes at least one of pages, characters, words, phrases, text-lines, sentences, paragraphs, columns of text, blocks of text, text articles, multi-page documents, collections of single-page documents and collections of multi-page documents.

[c7] The method of claim 1, wherein the OCR-output text passage triage decision includes at least one of sending the OCR-output text passage directly to an end user without post-OCR processing, sending the OCR-output text passage through a post-OCR inspection and processing stage, and sending the original text passage image to be keyed in manually.

[c8] The method of claim 1, wherein the triage model is a trained off-line triage model.

[c9] The method of claim 1, wherein the OCR-output text passage threshold error rate is a predetermined value.

[c10] The method of claim 7, wherein sending the OCR-output text passage through the post-OCR inspection and processing stage comprises:

determining at least one text passage error probability value for each OCR-output text passage as a correction operator detects and corrects an error in the OCR-output text passage; and

alerting the correction operator when the at least one text passage error probability value is improved so as to meet the OCR-output text passage threshold error value,

wherein the text passage error probability value for each OCR-output text passage is based on a probability of the at least one OCR-output character attribute being erroneously interpreted by the system.

[c11] The method of claim 10, wherein determining the text passage error probability value for an OCR-output text passage comprises:

determining OCR-output text passage error probability values for a plurality of selected portions of the OCR-output text passage; and arranging the plurality of selected portions of the OCR-output text passage based on the determined OCR-output text passage error probability values such that the selected portions having the highest OCR-output text passage error probability values are displayed first to the correction operator.

[c12] A computer-implemented method for triage of a plurality of OCR-output text passages, each OCR-output text passage having at least one OCR-output character, the method comprising:
selecting a set of OCR-output character attributes from a plurality of OCR-output character attributes for each OCR-output character;
determining an OCR-output character error value for each OCR-output character based on a probability of the set of OCR-output character attributes being erroneously interpreted by the OCR system;
determining a text passage error value for each OCR-output text passage based on a probability of the text passage being erroneously interpreted by the OCR system as determined using at least the OCR-output character error values; and
comparing the determined text passage error value with an OCR-output text passage threshold error value to perform an OCR-output text passage triage decision.

[c13] The computer-implemented method of claim 12, wherein the probability of the set of OCR-output character attributes being erroneously interpreted by the OCR system is determined based on at least the selected set of OCR-output character attributes processed using the triage model.

[c14] The computer-implemented method of claim 12, wherein the plurality of OCR-output character attributes includes at least one of a character class, a confidence descriptor class, a language of the text passage, a text passage publication date, a typeface in which the text passage is printed, an image-based feature of an individual character image and metadata attached to the text passage.

[c15] The computer-implemented method of claim 12, wherein the text passage to be triaged includes at least one of pages, characters, words, phrases, text-lines, sentences, paragraphs, columns of text, blocks of text, text articles, multi-page

documents, collections of single-page documents and collections of multi-page documents.

[c16] The computer-implemented method of claim 12, wherein the OCR-output text passage triage decision includes at least one of sending the OCR-output text passage directly to an end user without post-OCR processing, sending the OCR-output text passage through a post-OCR inspection and processing stage, and sending the original text passage image to be keyed in manually.

[c17] The computer-implemented method of claim 16, wherein sending the OCR-output text passage through a post-OCR inspection and processing stage comprises:
determining at least one text passage error probability value for each OCR-output text passage as a correction operator detects and corrects an error in the OCR-output text passage; and
alerting the correction operator when the at least one text passage error probability value is improved so as to meet the OCR-output text passage threshold error value,
wherein the text passage error probability value for each OCR-output text passage is based on a probability of the at least one OCR-output character attribute being erroneously interpreted by the system.

[c18] The computer-implemented method of claim 12, wherein determining a text passage error probability value for an OCR-output text passage comprises:
determining OCR-output text passage error probability values for a plurality of selected portions of the OCR-output text passage; and
arranging the plurality of selected portions of the OCR-output text passage based on the determined OCR-output text passage error probability values such that the selected portions having the highest OCR-output text passage error probability values are displayed first to the correction operator.

[c19] An OCR-output text passage triage system that triages a text passage outputted by an optical character recognition system, the OCR-output text passage including at least one OCR-output character having at least one OCR-output character attribute, the system comprising:

an OCR-output text passage character accuracy determination circuit or routine that determines a character interpretation error value using a triage model;
an OCR-output text passage accuracy determination circuit or routine that determines at least one OCR-output text passage quality metric using the determined character interpretation error value and at least one statistical algorithm or model included in the triage model; and
an OCR-output text passage triage circuit or routine that performs one or more text passage triage decisions using the determined at least one OCR-output text passage quality metric and an OCR-output text passage threshold error rate value.

[c20] The OCR-output text passage triage system of claim 19, wherein the triage model is a trained off-line triage model.

[c21] The OCR-output text passage triage system of claim 19, wherein the OCR-output text passage threshold error rate value is included in a text passage error threshold operating point model.

[c22] The OCR-output text passage triage system of claim 19, wherein the at least one OCR-output character attribute includes at least one of a character class, a confidence descriptor class, a language of the text passage, a text passage publication date, a typeface in which the text passage is printed, an image-based feature of an individual character image and metadata attached to the text passage.

[c23] The OCR-output text passage triage system of claim 19, wherein the text passage to be triaged includes at least one of pages, characters, words, phrases, text-lines, sentences, paragraphs, columns of text, blocks of text, text articles, multi-page documents, collections of single-page documents and collections of multi-page documents.

[c24] The OCR-output text passage triage system of claim 19, wherein the OCR-output text passage triage decision includes at least one of sending the OCR-output text passage directly to an end user without post-OCR rekeying or correction, sending the OCR-output text passage through a post-OCR

inspection and correction stage, and sending the original text passage image to be completely keyed in manually.

[c25] A machine-readable medium that provides instructions for triage of a text passage outputted by an optical character recognition system, the OCR-output text passage having at least one OCR-output character, instructions, which when executed by a processor, cause the processor to perform operations comprising:
determining at least one OCR-output character attribute for each OCR-output character;
determining an error rate for the OCR-output text passage using a triage model and the determined at least one OCR-output character attribute; and
comparing the determined error rate for the OCR-output text passage with an OCR-output text passage threshold error rate to perform an OCR-output text passage triage decision.

[c26] The machine-readable medium of claim 25, wherein determining an error rate for the OCR-output text passage comprises:
providing the at least one OCR-output character attribute to the triage model;
determining a character interpretation error value for each OCR-output character based on a probability of the at least one OCR-output character attribute being erroneously interpreted by the system; and
determining a text passage error value based on the at least one character interpretation error value determined for each OCR-output character.

[c27] The machine-readable medium of claim 26, further comprising:
determining a number representing a sum of OCR-output characters in the OCR-output text passage; and
dividing the text passage error value by the number representing the sum of OCR-output characters.

[c28] The machine-readable medium of claim 25, wherein determining at least one OCR-output character attribute for each OCR-output character comprises selecting the at least one OCR-output character attribute from a plurality of OCR-output character attributes.

[c29] The machine-readable medium of claim 28, wherein the plurality of OCR-output character attributes includes at least one of a character class, a confidence descriptor class, a language of the text passage, a text passage publication date, a typeface in which the text passage is printed, an image-based feature of an individual character image and metadata attached to the text passage.

[c30] The machine-readable medium of claim 25, wherein the text passage to be triaged includes at least one of pages, characters, words, phrases, text-lines, sentences, paragraphs, columns of text, blocks of text, text articles, multi-page documents, collections of single-page documents and collections of multi-page documents.

[c31] The machine-readable medium of claim 25, wherein the OCR-output text passage triage decision includes at least one of sending the OCR-output text passage directly to an end user without post-OCR processing, sending the OCR-output text passage through a post-OCR inspection and processing stage, and sending the original text passage image to be keyed in manually.